Jestr Journal of Engineering Science and Technology Review 8 (5) (2015) 110- 116

JOURNAL OF Engineering Science and Technology Review

Research Article

www.jestr.org

Spatial Outlier Detection of CO_2 Monitoring Data Based on Spatial Local Outlier Factor

Liu Xin [1, 2], Zhang Shaoliang [1,*] and Zheng Pulin [3]

[1]*School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou, Jiangsu 221008, China*
[2]*School of Medicine Information of Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*
[3] *The University of Central Queensland, 44 Greenhill Road, QLD, Australia*

Received 14 August 2015; Accepted 15 December 2015

Abstract

Spatial local outlier factor (SLOF) algorithm was adopted in this study for spatial outlier detection because of the limitations of the traditional static threshold detection. Based on the spatial characteristics of CO_2 monitoring data obtained in the carbon capture and storage (CCS) project, the K-Nearest Neighbour (KNN) graph was constructed using the latitude and longitude information of the monitoring points to identify the spatial neighbourhood of the monitoring points. Then SLOF was adopted to calculate the outlier degrees of the monitoring points and the 3σ rule was employed to identify the spatial outlier. Finally, the selection of K value was analysed and the optimal one was selected. The results show that, compared with the static threshold method, the proposed algorithm has a higher detection precision. It can overcome the shortcomings of the static threshold method and improve the accuracy and diversity of local outlier detection, which provides a reliable reference for the safety assessment and warning of CCS monitoring.

Keywords: Spatial local outlier factor, Carbon capture and storage, CO_2, Spatial outlier

1. Introduction

Global warming caused by the increasing concentration of CO_2 has become a hot topic in the 21st century [1]. Carbon capture and storage (CCS) technology has been recognized as an effective strategy to lower carbon dioxide emissions directly [2]. Based on the estimates of the International Energy Agency (IEA), the contribution of the CCS technique to the global CO_2 reduction will reach 19% by 2050 [3]. However, based on the survey on the recognition and acceptance of the public of CCS projects, the sudden large-scale leakage of CO_2 is the main risk for CO_2 geological storage. Underground CO_2 can leak to the surface through cracks, abandoned wells, and other channels (Fig. 1). Therefore, to ensure the safety of CCS projects, adopting effective monitoring measures and identifying leakage warning signs in a timely manner are necessary to provide the basis for emergency rescue.

Outlier detection is conducive to identifying warning signs and providing the basis for warning. CCS monitoring data has various forms (i.e., time series data, spatial data, seismic data, etc.). The monitoring data of surface CO_2 concentration are taken as an example for outlier detection in this study. Because CO_2 concentration monitoring data belong to the spatial data with significant spatial characteristics, the traditional threshold method ignores the spatial characteristic of the data and fails to effectively identify local outlier caused by leakage. Thus, considering the spatial characteristics of the data [5], the spatial local outlier factor (SLOF) algorithm is adopted for the outlier detection of CO_2 concentration monitoring data, hoping that the relevant results can provide a reliable reference for safety assessment and warning of CCS monitoring.

2. State of the Art

The original spatial outlier mining algorithms were based on the idea of spatial statistics with variation image and scatter chart Z-value algorithms [6] as representatives. Those algorithms failed to consider the characteristics of spatial data and demonstrated poor mining effect. Shekhar et al. [7] first proposed the algorithm that distinguished spatial and non-spatial attributes, obtained the spatial neighborhood through building the adjacency relation between entities (or building the KNN relationship), and identified the spatial outlier through the difference in non-spatial attributes between spatial entity and its neighboring entity. However, this method was only suitable for identifying global outliers. Breunig et al. [8] proposed the local outlier factor (LOF) concept, which was a density-based method to identify outlier points through the local outlier degree. With clear physical significance, this method could effectively mine global or local outlier points. The local correlation integral (LOCI) method was also presented to discover local outliers [9]. However, it did not distinguish the spatial and non-spatial attributes, resulting in the difficulty in explaining the

* E-mail address: flyinsky6@189.cn
ISSN: 1791-2377

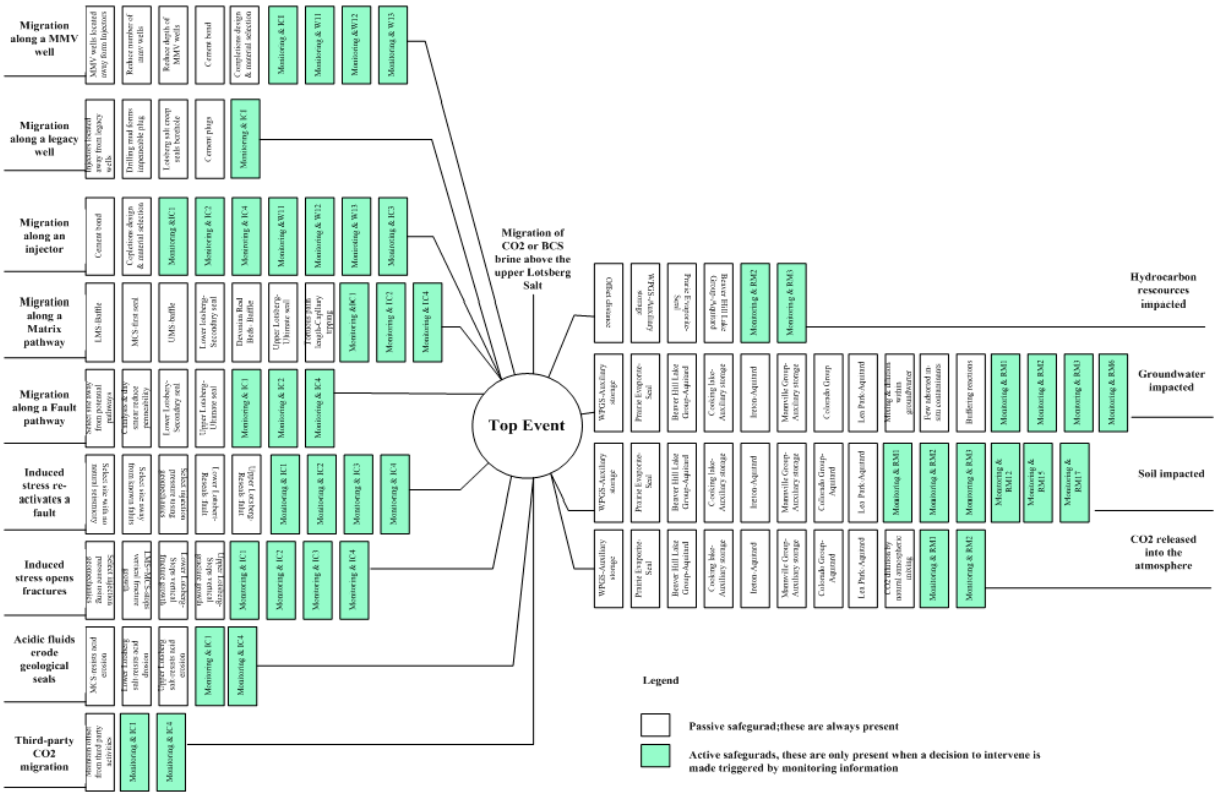

Fig.1. Possible escape passage and risk map of CO_2 geological storage [4]

detection results [10].

Therefore, numerous scholars consolidated the SLZ and LOF algorithms to propose the spatial local outlier measure (SLOM) [11] and spatial local outlier factor [12], among others. In the SLOM algorithm [13], the fluctuating factor is determined only by symmetrical distribution. In the cases of few spatial neighbors or slight fluctuations, showing the fluctuation is difficult, which results in high missing and error detection rates [14]. The SLOF algorithm is used to construct the neighborhood with spatial attributes and develop an outlier measure with non-spatial attributes based on the LOF algorithm. This algorithm demonstrates high precision, low user dependence, and high operation efficiency. CO_2 concentration monitoring data belong to the spatial data that have significant spatial autocorrelation [15] and spatial heterogeneity [16]. Thus, the SLOF algorithm is selected for CO_2 outlier detection.

3. Spatial Outlier Detection of CO2 Monitoring Data Based on SLOF

The process of spatial outlier detection is divided into spatial outlier measure and identification [17]. The neighborhood in this study is constructed based on the spatial attributes of the monitoring data, and the local outlier coefficient of the monitoring points is calculated by the SLOF algorithm to determine the outliers identified through the 3-D graph. The specific steps are as follows:

If N monitoring points exist, they constitute the object set $o=\{o_1,o_2,...,o_n\}$. The spatial attribute of Object $o\in O$ is written as $(lat_o,\log_o)$, and the d dimensional spatial attribute is written as $f(o)$, expressed as $(f(o_1),f(o_2),...,f(o_d))$ σ_c which represents the spatial adjacency relation under the specified condition C.

3.1 Construction of the Spatial Neighborhood

The spatial neighborhood can be constructed in three ways: ε adjacent connection, K-nearest neighbor, and full connection graphs [18]. In this study, the most commonly used K-nearest neighbor connection graph is adopted to construct the spatial neighborhood. The specific construction process is as follows:

The latitude and longitude of a monitoring point O_i are $(lat_i,\log_i)$ respectively. The Euclidean distance from other monitoring points is known as the nearest neighbor distance, written as $dist(i,j)$. The calculation formula is as follows:

$$dist(i,j)=\sqrt{(\log_i-\log_j)^2+(lat_i-lat_j)^2}\quad j=1,2...,n \qquad (1)$$

The KNN graph connects O_i with its K neighbors by selecting the K-nearest neighbor nodes before O_i to form the K-directed edges of the KNN graph. However, the directed graph constructed in this manner does not meet the symmetry requirement of graph theory. Thus, the following method is adopted to solve this problem: if O_i is one of the K neighbors of O_j or O_j is one of the K neighbors of O_i, then a connection exists between O_i and O_j. The graph obtained in this manner is known as the K-nearest neighbor graph.

Definition 1 Spatial Neighbor

For the spatial object O_i, the objects that have a connection relation with O_i in the KNN graph are known as the spatial neighbors of O_i.

Definition 2 Spatial Neighborhood

All spatial neighbors constitute the spatial neighborhood of O_i, written as $NB(o_i)=\{o_1,o_2,..,o_m\}$, ($m \geq k$).

3.2 Spatial Outlier Measure

The spatial outlier measure of Object O ready to be detected is obtained by calculating the ratio of the neighborhood distance of 0 to the average neighborhood distance of its neighbors. To eliminate the difference in the dimensions of different monitoring parameters, the initial data are normalized before calculating the neighborhood distance to align them into a specific area [0, 1]. The normalization formula is as follows:

$$x' = \frac{x_{\max} - x}{x_{\max} - x_{\min}} \tag{2}$$

Where x' represents the data after normalization, x represents the initial data, and $x_{\max}, x_{\min}$ represent the maximum and minimum values of the data respectively.

Definition 3 Neighborhood Distance between Objects

The neighborhood distance between P and O is calculated with the weighted Euclidean distance. The formula is as follows:

$$dist(p,o,w)=\sqrt{\sum_{k=1}^{d} w_k (f(o_k)-f(p_k))^2} \tag{3}$$

Where w_k is the weight of non-spatial attributes; $\sum_{k=1}^{d} w_k = 1$, d is the number of non-spatial attributes; and $f(o_k)$ is the k th non-spatial attribute value after the normalization of O. The neighborhood distance between objects represents the dissimilarity between the object and its neighborhood in non-spatial attribute. The greater the neighborhood distance, the higher the dissimilarity.

Definition 4 Neighborhood Distance of Object O

The neighborhood distance of Object O refers to the average of the weighted distances between O and all objects in its spatial neighborhood. The formula is as follows:

$$Ndist(o,w)=\frac{\sum_{p\in NB(o)} dist(p,o,w)}{|NB(o)|} \tag{4}$$

Where $|NB(o)|$ refers to the number of spatial neighbors of Object O. Given that all non-spatial attributes are processed by normalization and $\sum_{k=1}^{d} w_k = 1$, then $0 \leq Ndist(o,w) \leq 1$.

Definition 5 Spatial Local Outlier Coefficient of Object O

The spatial local outlier coefficient of Object O refers to the ratio of the neighborhood distance of O to the average neighborhood distance between 0 and its neighbors. This coefficient can be expressed by $SLOF(o)$. The formula is as follows:

$$SLOF(o)=\frac{Ndist(o,w)}{\frac{\sum_{p\in NB(o)} Ndist(p)}{|NB(o)|}} \tag{5}$$

$SLOF(o)$ represents the local outlier degree of O. The greater the $SLOF$, the higher the outlier degree of the object.

3.3 Identification of Spatial Outlier

Spatial outlier identification is easily ignored in numerous spatial outlier detections. The $SLOF$ values of all objects are generally sorted in a descending order. The first n objects with the highest outlier degree are the spatial outlier points in question. This outlier detection method is subjective and random without theoretical support. Based on the analysis of the time-variant characteristics of CO_2 monitoring data and their distribution, in line with the Chebyshev law of large numbers and central limit theorem, CO_2 monitoring data are considered to obey the normal distribution (proof omitted). The 3σ rule can be selected as the basis for spatial outlier identification.

3σ Rule: If $X \sim N(\mu,\sigma^2)$, the probability that the normal observation values should distribute between $(\mu-3\sigma,\mu+3\sigma)$ is 99.74%, where μ is the average value in the window and σ is the standard deviation of the data in it.

The outliers identified by the 3σ rule are determined based on prior knowledge.

4. Example Analysis

The CO_2 monitoring data of a storage area in August were considered for outlier detection and performance analysis in this study. The algorithms were realized in Mat lab R2014a. The operation environment was Win7, Intel(R) Core(TM) i5, CPU3.2 GHz, 4 GB RAM.

4.1 Experimental Data

CO_2 monitoring data were composed of two spatial attributes (longitude and latitude) and six non-spatial attributes (CO_2 concentration, temperature, humidity, wind speed, pressure, and altitude). Some data with empty monitoring records were deleted through data preprocessing. Finally, 73 monitoring points were selected. The following figure shows the distribution map of the monitoring points.

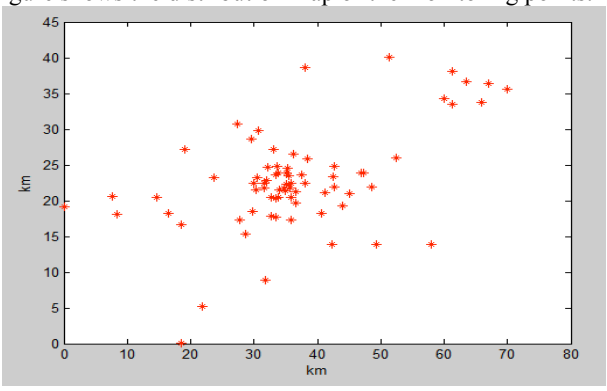

Fig. 2. Distribution map of monitoring points

The monitoring values were modified to add the outlier points and compare the proposed algorithm with the static threshold method. In the static threshold method, 1000 ppm

was set as the threshold of outlier detection. Thus, three monitoring points in this study were modified to be less than1000 ppm, in which the value of the No. 19 monitoring point was modified from 393 ppm to 980 ppm, No. 36 from 416 ppm to 870 ppm, and No. 50 from 395 ppm to 990 ppm.The boxplot of six detecting parameters was drawn, as shown in Figure 3.

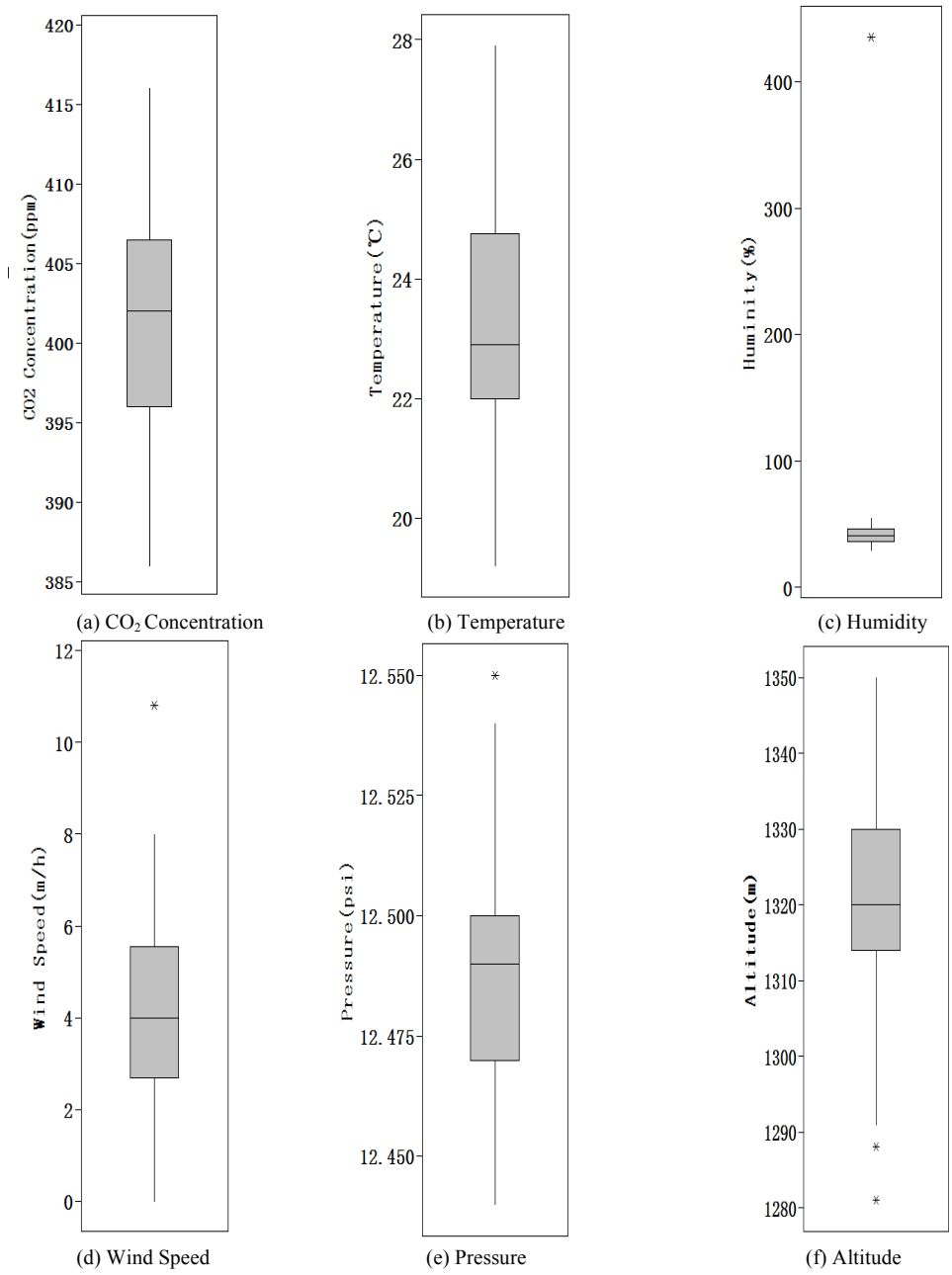

Fig. 3. Boxplot of CO2 monitoring parameters

Figure 3 shows that the medians of CO_2 concentration, temperature, humidity, wind speed, pressure, and altitude are 400 ppm, 22.6 °C, 41%, 4 m/h, 12.49 psi, and 1320 m, respectively. At the same time, the transverse line in the middle of the CO_2 concentration box is located at the lower position of the box, indicating the existence of the outlier value. The monitoring data can be analyzed through the extremum, which belongs to the category of outlier point detection. This study focuses on the research of local outliers because of space limitations.

4.2 Analysis of Detection Results

The spatial local outlier coefficients of the monitoring points were calculated using Formulas (2) to (5). The first five maximum values were selected, as shown in Table 1:

Tab. 1. Spatial Local Outlier Coefficient (K = 5)

No.	Monitoring Point No.	Spatial Local Outlier Coefficient
1	50	6.08
2	19	5.63
3	36	4.85
4	27	1.75
5	56	1.66

CO_2 concentration is the most intuitive reflection of leakage, and thus the accuracy of the algorithm can be determined based on the CO_2 concentration outlier. The following figure presents the 3-D graphs of the longitudes, latitudes, and CO_2 concentrations of the first five spatial outlier points. The circles in the graphs represent the neighbor distribution of the outlier points in the neighborhood.

Figure 4 are the scatter plots, the x, y axis are the km net coordinates, unit is km, z axis is the CO_2 concentration, and unit is ppm. The first three local outlier points are all global; the concentration value of the fourth outlier point was 396 ppm, slightly lower than those of the surrounding outlier points with values between 398 and 414 ppm. The concentration value of the fifth outlier point is 415 ppm, slightly higher than those of the surrounding outlier points with the values between 388 and 404 ppm. Therefore, the first five outlier points identified by SLOF are all correct.

The spatial outlier was identified by the 3σ rule. After calculation, the average of the outlier degree was 0.93, the mean square was 1.029, and the identified outlier points were 19, 36, and 50.

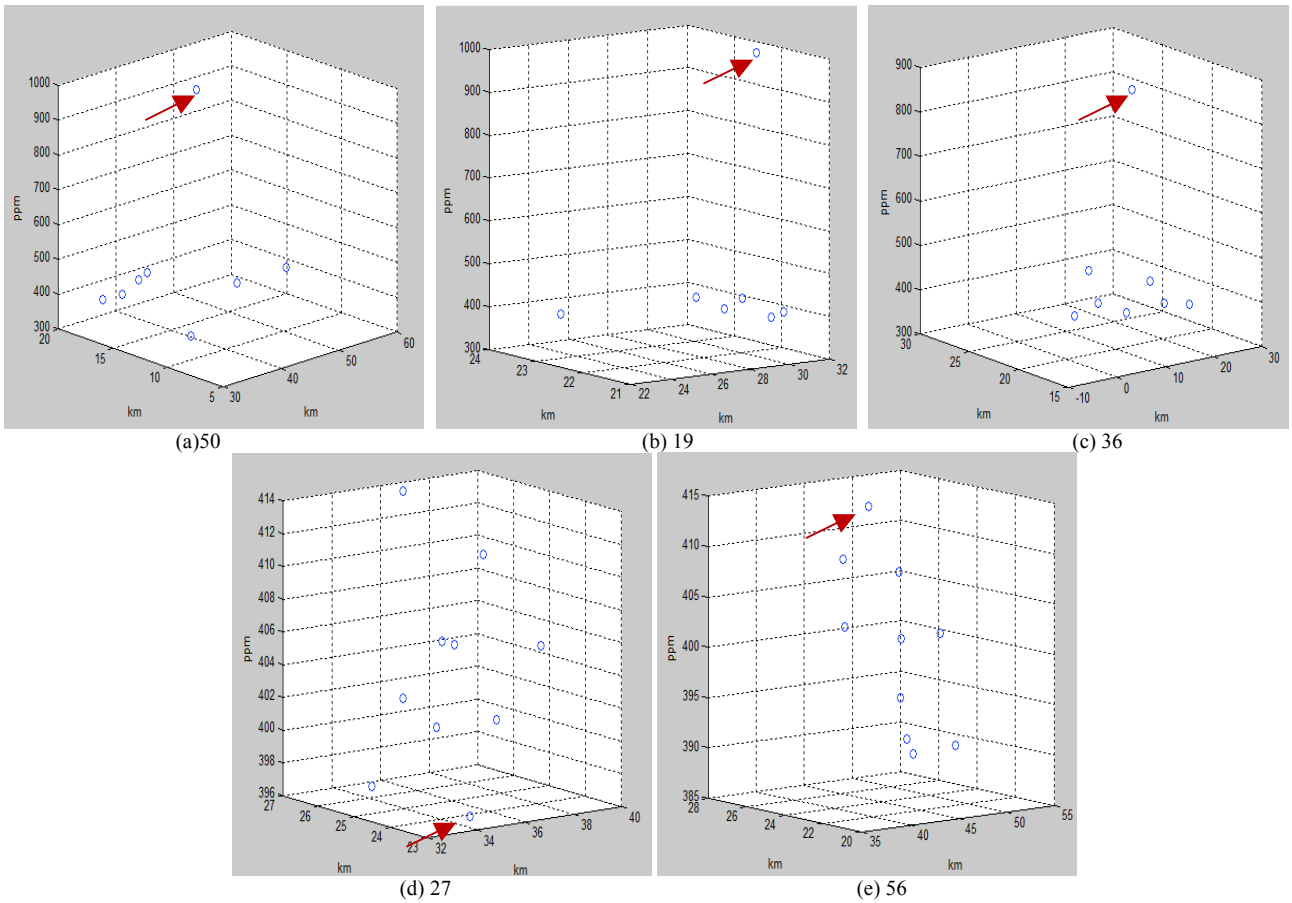

Fig. 4. Scatter plot of the first five local outlier points

Based on the static threshold method, if the threshold value is set to 1000 ppm, then no outlier point can be identified, which easily causes missing detection. Thus, *SLOF* and 3σ can improve the precision of outlier detection and automatically identify the outlier points to avoid the randomness and uncertainty of traditional outlier identification.

4.3 Analysis of the Influence of K Value on Outlier Point Detection

From the above mentioned research, the SLOF algorithm shows significant improvement in detection precision compared with the traditional static threshold method. However, the selection of K value is the main problem of SLOF. On one hand, a too large or small K value reduces the precision of outlier detection; on the other hand, K value affects the time complexity of the algorithm. Taking the construction of the KNN graph as an example. The time complexity of this graph can reach O (KN^2), in which K is the number of neighbors and N is the total number of data points. Thus, the selection of K value in this study is analyzed to finally determine the suitable K value for this group of data. Based on the analysis, we decide to analyze the outlier point detection when K < 5.The calculated spatial local outlier coefficients when K = 4 are shown in Table 2:

Tab. 2. Spatial Local Outlier Coefficient ($K = 4$)

No.	Monitoring Point No.	Spatial Local Outlier Coefficient
1	19	5.35
2	50	4.76
3	36	4.396
4	27	1.76
5	2	1.736

After calculation, the mean value of the outlier degree is 0.94 and the mean square is 0.397. Three outlier points are identified based on the 3σ rule, namely, 19, 36, and50.

The calculated spatial local outlier coefficients when K = 3 are shown in Table 3:

Tab. 3. Spatial Local Outlier Coefficient (K = 3)

No.	Monitoring Point No.	Spatial Local Outlier Coefficient
1	19	3.56
2	50	3.49
3	36	3.14
4	27	1.63
5	2	1.57

After calculation, the mean value of the outlier degree is 0.93 and the mean square is 0.198. Fifteen outlier points are

identified based on the rule, namely, 3, 8, 10, 11, 19, 27, 36, 40, 50, 51, 58, 61, 63, 66, and 67.

The detection and false alarm rates, which are commonly used in outlier detection, are adopted as the indicators that measure the performance of the algorithm to analyze the influence of K value on outlier point detection. The specific calculation process is as follows:

*Detection rate = (the number of samples detected as outlier event / all samples of outlier events) * 100%*
*False alarm rate = (the number of normal samples detected as outlier events / all normal samples) * 100%*

The Figure 5 presents the graphs of the detection and false alarm rates when K = 3, 4 and 5.

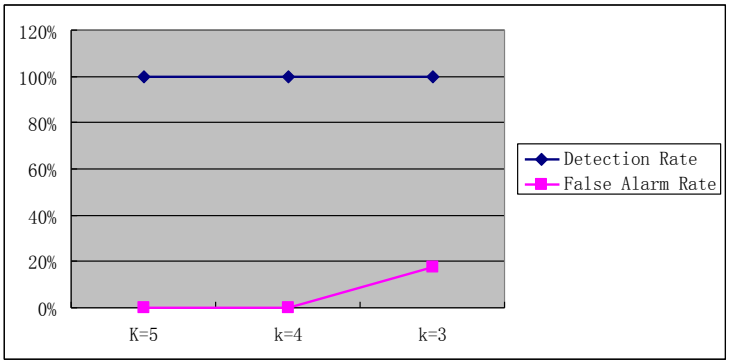

Fig. 5. Graphs of detection and false alarm rates under different K values

Based on figure 5, when K > 4, the algorithm can achieve 100% detection rate and 0% false alarm rate; however, when K = 3, the false alarm rate increases to 17%. Thus, based on the detection accuracy and time complexity, the optimal K value is 4 in the outlier detection of the data.

5. Conclusion

Given the shortcomings of the traditional static threshold method, the present status of the research on spatial outlier detection was analyzed in this study. Based on the characteristics of CCS monitoring data, the SLOF algorithm was adopted for spatial outlier detection of CCS monitoring data. First of all, this algorithm constructed the KNN graph based on the spatial information of the monitoring points to determine the spatial neighborhood of all monitoring points. The weighted Euclidean distance of the non-spatial attributes of the monitoring data was then used to calculate the neighborhood distances between objects. In addition, the SLOF algorithm was adopted to calculate the spatial local outlier coefficients of objects. Finally, the 3σ rule that could automatically identify the spatial outlier points was adopted for local outlier detection. The CO_2 concentration monitoring data in August in a year with additional outlier points were used for local spatial outlier identification to verify the accuracy of the algorithm. The detection and false alarm rates under different K values were analyzed to select the optimal one in the experimental data.

From the research, the SLOF algorithm can effectively detect the spatial outlier compared with the static threshold method. However, the selection of K value has some uncertainty, which will affect the detection effect. Therefore, to further improve the detection precision, the research on other neighborhood construction methods can be considered to avoid the effect of parameter uncertainty on the detection result.

Acknowledgements

The paper was supported by the Project (No. 2011BAC08B03) of The 12th Five Year science and technology support Plan, and the Project (No. SZBF2011-6-B35) of the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Yan J.Y., "Carbon Capture and Storage (CCS)". *Applied Energy*,148, 2015, pp.A1-A6.
2. Zhu L., Duan H.B, Fan Y., "CO_2 mitigation potential of CCS in China-an evaluation based on an integrated assessment model". *Journal of Cleaner production*, 103, 2015, pp.934-947.
3. Stephen W., Blackford J.C., John I.S., "Assessing the environmental consequences of CO_2 leakage from geological CCS: Generating evidence to support environmental risk assessment". *Marine Pollution Bulletin*, 73(2), 2013, pp.399-401.
4. Bourne S., Crouch S., Smith M., "A risk-based framework for measurement, monitoring and verification of the Quest CCS Project, Alberta, Canada". *International Journal of Greenhouse Gas Control*, 26, 2014, pp.109-126.
5. Chongcheng C., Jiaxiang J., Xiaozhu W., "Parallel and Distributed Spatial Outlier Mining in Grid: Algorithm, Design and Application". *Journal of Grid Computer*, 13(2), 2015, pp.139-157.
6. Liu X.T., Chen F., Lu C.T, "On detecting spatial categorical outliers". *Geoinformatica*, 18, 2014, pp.501-536.
7. Shekhar S., Lu C., Zhang P., "A Unified Approach to Detecting Spatial Outliers".*GeoInformatica*,7(2),2003,pp.139 - 166.
8. Markus M. Breunig, Hans-Peter Kriegel, Raymond T.Ng, Jorg Sander, "LOF: Identifying Density-based Local Outliers". *Proceedings of the International Conference on Management of Data*, Dallas, TX USA, May, 2000, pp.93-104.
9. Papadimitriou S., Kitagawa H., Gibbons P., Faloutso C., "Loci: fast outlier detection using the local correlation integral", *Proceedings of the 19th International Conference on Data Engineering*, Bangalore, India, March, 2003, pp.315-328.
10. Grekousis G. Fotis Y.N., "A fuzzy index for detecting spatiotemporal outliers", *Geoinformatica*, 16(3), 2012, pp.597-619.
11. Chawla S., Sun P., "SLOM: a new measure for local spatial outliers", *Knowledge and Information Systems,* 9(4), 2006, pp.412-429.
12. Xue A.R., "Study on Spatial Outlier Mining", Zhenjiang, Doctoral Dissertation, China.
13. Janeja, V.P., Palanisamy R., "Multi-domain Anomaly Detection in Spatial Datasets". *Knowledge and Information Systems*, 36(3), 2013, pp.749-788.
14. Chen C.C, Lin J.X., Wu X.Z., "Parallel and Distributed Spatial Outlier Mining in Grid: Algorithm, Design and Application", *Journal of grid computing*, 13(2), 2015, pp.139-157.

15. Alvera-Azcarate A., Sirjacobs D., Barth A., "Outlier detection in satellite data using spatial coherence". *Remote Sensing of Environment*, 119,2012, pp.84-91.
16. Cai Q., He H.B., Man H., "Spatial outlier detection based on iterative self-organizing learning model", *Neurocomputing*, 117, 2013, pp.161-172.
17. Deng M., Liu Q.L., Li G.Q., "Spatial outlier detection method based on spatial clustering", *Journal of Remote Sensing*,14(5), 2010, pp.944-958.
18. Jia J.H., "*Research on Spectral clustering integration algorithm*", Tianjin: Tianjin university press , China, 2011.